

Provide distinct cgroup VM overcommit accounting and handling using the memory resource controller.

Patchset against latest Linus git tree.

This patchset allows to set different per-cgroup overcommit rules and, according to them, it's possible to return a memory allocation failure (ENOMEM) to the applications, instead of always triggering the OOM killer via `mem_cgroup_out_of_memory()` when cgroup memory limits are exceeded.

Default overcommit settings are taken from `vm.overcommit_memory` and `vm.overcommit_ratio` sysctl values. Child cgroups initially inherits the VM overcommit parent's settings.

Cgroup overcommit settings can be overridden using `memory.overcommit_memory` and `memory.overcommit_ratio` files under the cgroup filesystem.

For example:

1. Initialize a cgroup with 50MB memory limit:

```
# mount -t cgroup none /cgroups -o memory  
# mkdir /cgroups/0  
# /bin/echo $$ > /cgroups/0/tasks  
# /bin/echo 50M > /cgroups/0/memory.limit_in_bytes
```
2. Use the "never overcommit" policy with 50% ratio:

```
# /bin/echo 2 > /cgroups/0/memory.overcommit_memory  
# /bin/echo 50 > /cgroups/0/memory.overcommit_ratio
```

Assuming we have no swap space, cgroup 0 can allocate up to 25MB of virtual memory. If that limit is exceeded all the further allocation attempts made by userspace applications will receive a -ENOMEM.

4. Show committed VM statistics:

```
# cat /cgroups/0/memory.overcommit_as  
CommitLimit: 25600 kB  
Committed_AS: 9844 kB
```

5. Use "always overcommit":

```
# /bin/echo 1 > /cgroups/0/memory.overcommit_memory
```

This is very similar to the default memory controller configuration: overcommit is allowed, but when there's no more available memory oom-killer is invoked.

TODO:

- shared memory is not taken in account (i.e. files in tmpfs)

-Andrea

Containers mailing list

Containers@lists.linux-foundation.org

<https://lists.linux-foundation.org/mailman/listinfo/containers>
